Github2023

Keely Grice

12/20/2023

```
library('tidyverse')
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library('stargazer')
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library('sandwich')
```

```
## Warning: package 'sandwich' was built under R version 4.3.2
```

WHI Data

```
whi <- read.csv('whidata.csv')
#Making a Balance Table
trt1 <- subset(whi, treat == 1)
con1 <- subset(whi, treat == 0)

trt1 <- trt1 %>%
  group_by(age) %>%
  summarise(n())

con1 <- con1 %>%
  group_by(age) %>%
  summarise(n())

treat_df <- data.frame(
  Group = c("50-59", "60-69", "70-79"),
  Treatment = trt1$`n()`,
  Control=con1$`n()`)

stargazer(t(treat_df),
          type = "text",
          title = "Comparison for Treatment by Age",
          digits = 2,
          summary = FALSE)
```

```
##
## Comparison for Treatment by Age
## ============================
## Group     50-59 60-69 70-79
## Treatment 2837  3854  1815
## Control   2683  3655  1764
## ----------------------------
```

```
#Data Analysis
mean(whi$breastcancer[whi$treat==1]) - mean(whi$breastcancer[whi$treat==0])
```

```
## [1] 0.00508712
```

```
#Treated women have a .51% higher incidence of breast cancer than untreated women in this study.

#Confidence Interval for Bootstrapped Mean Difference
set.seed(122)
n_ite <- 1000
tau_treat <- rep(NA, n_ite)
n_sample <- nrow(whi)

for(i in 1:n_ite){
  index_bs <- sample(1:n_sample, replace = T)
  df_bs <- whi[index_bs, ]
  tau_treat[i] <- mean(df_bs$breastcancer[df_bs$treat==1])-mean(df_bs$breastcancer[df_bs$treat==
0])}
mean(tau_treat)
```

```
## [1] 0.005066242
```

```
quantile(tau_treat, c(0.025, 0.975))
```

```
##        2.5%       97.5%
## 0.000375398 0.009337439
```

```
#Treatement Effect Based on Age
low_age <- whi %>%
  filter(age=="50-59")
mean(low_age$breastcancer[low_age$treat==1]) - mean(low_age$breastcancer[low_age$treat==0])
```

```
## [1] 0.003732558
```

```
med_age <- whi %>%
  filter(age=='60-69')
mean(med_age$breastcancer[med_age$treat==1]) - mean(med_age$breastcancer[med_age$treat==0])
```

```
## [1] 0.00464882
```

```
high_age <- whi %>%
  filter(age=='70-79')
mean(high_age$breastcancer[high_age$treat==1]) - mean(high_age$breastcancer[high_age$treat==0])
```

```
## [1] 0.008210116
```

```
#Bootstrapped Treatment Effect Based on Age
set.seed(679)
n_ite <- 1000
tau_treat1 <- rep(NA, n_ite)
n_sample <- nrow(whi)
tau_fifty=c()
tau_sixty=c()
tau_seventy=c()

for(i in 1:n_ite){
  index_bs <- sample(1:n_sample, replace=T)
  df_bs<- whi[index_bs, ]
  if (df_bs$age[i] == '50-59'){
    a <- mean(df_bs$breastcancer[df_bs$treat ==1])-mean(df_bs$breastcancer[df_bs$treat==0])
    tau_fifty <- append(tau_fifty, a)
  }
  else if (df_bs$age[i] == '60-69'){
    b <- mean(df_bs$breastcancer[df_bs$treat ==1])-mean(df_bs$breastcancer[df_bs$treat==0])
    tau_sixty <- append(tau_sixty, b)
  }
  else {
    c <- mean(df_bs$breastcancer[df_bs$treat ==1])-mean(df_bs$breastcancer[df_bs$treat==0])
    tau_seventy <- append(tau_seventy, c)
  }
}

quantile(tau_fifty, c(0.025, 0.975))
```

```
##         2.5%        97.5%
## 0.0007806639 0.0095115745
```

```
quantile(tau_sixty, c(0.025, 0.975))
```

```
##         2.5%        97.5%
## 0.0007972799 0.0095311819
```

```
quantile(tau_seventy, c(0.025, 0.975))
```

```
##        2.5%       97.5%
## 0.001137086 0.008701243
```

```
#Regression Analysis
whi <- whi %>% mutate(fifties=ifelse(whi$age == '50-59', 1, 0),
                      sixties = ifelse(whi$age == '60-69', 1, 0),
                      seventies = ifelse(whi$age == '70-79', 1, 0)
)

#Additive Regression
lm_whi <- lm(breastcancer~treat + sixties+ seventies, data=whi)

#Reference category: 50-59
mean(low_age$breastcancer[low_age$treat==0])
```

```
## [1] 0.01565412
```

```
stargazer(lm_whi,
          covariate.labels = c("Treatment", "60-69", "70-79"),
          title = "Breast Cancer", type = "text")
```

```
##
## Breast Cancer
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                            breastcancer
## ------------------------------------------------
## Treatment                     0.005**
##                               (0.002)
##
## 60-69                         0.005**
##                               (0.003)
##
## 70-79                        0.008***
##                               (0.003)
##
## Constant                     0.015***
##                               (0.002)
##
## ------------------------------------------------
## Observations                  16,608
## R2                             0.001
## Adjusted R2                    0.001
## Residual Std. Error     0.146 (df = 16604)
## F Statistic          4.247*** (df = 3; 16604)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
#Interactive Regression
lm_whi2 <- lm(breastcancer~treat + sixties+ seventies + treat*sixties + treat*seventies, data=wh
i)

stargazer(lm_whi, lm_whi2,
          covariate.labels = c("Treatment", "60-69", "70-79", "Treatment:60-69", "Treatment:70-7
9"),
          title = "Breast Cancer", type = "text")
```

```
##
## Breast Cancer
## ========================================================================
##                                    Dependent variable:
##                     -----------------------------------------------
##                                         breastcancer
##                                (1)                      (2)
## ------------------------------------------------------------------------
## Treatment                    0.005**                   0.004
##                              (0.002)                  (0.004)
##
## 60-69                        0.005**                   0.005
##                              (0.003)                  (0.004)
##
## 70-79                        0.008***                  0.006
##                              (0.003)                  (0.004)
##
## Treatment:60-69                                        0.001
##                                                       (0.005)
##
## Treatment:70-79                                        0.004
##                                                       (0.006)
##
## Constant                     0.015***                0.016***
##                              (0.002)                  (0.003)
##
## ------------------------------------------------------------------------
## Observations                  16,608                   16,608
## R2                            0.001                    0.001
## Adjusted R2                   0.001                    0.0005
## Residual Std. Error    0.146 (df = 16604)      0.146 (df = 16602)
## F Statistic         4.247*** (df = 3; 16604) 2.657** (df = 5; 16602)
## ========================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

The calculated mean differences are extremely similar to the interactive regression differences. However, the additive is less similar, especially for the age group 70-79 where it underestimates considerably. The interactive regression seems to be a better calculation for the average treatment effect than the additive.

California CPS Data

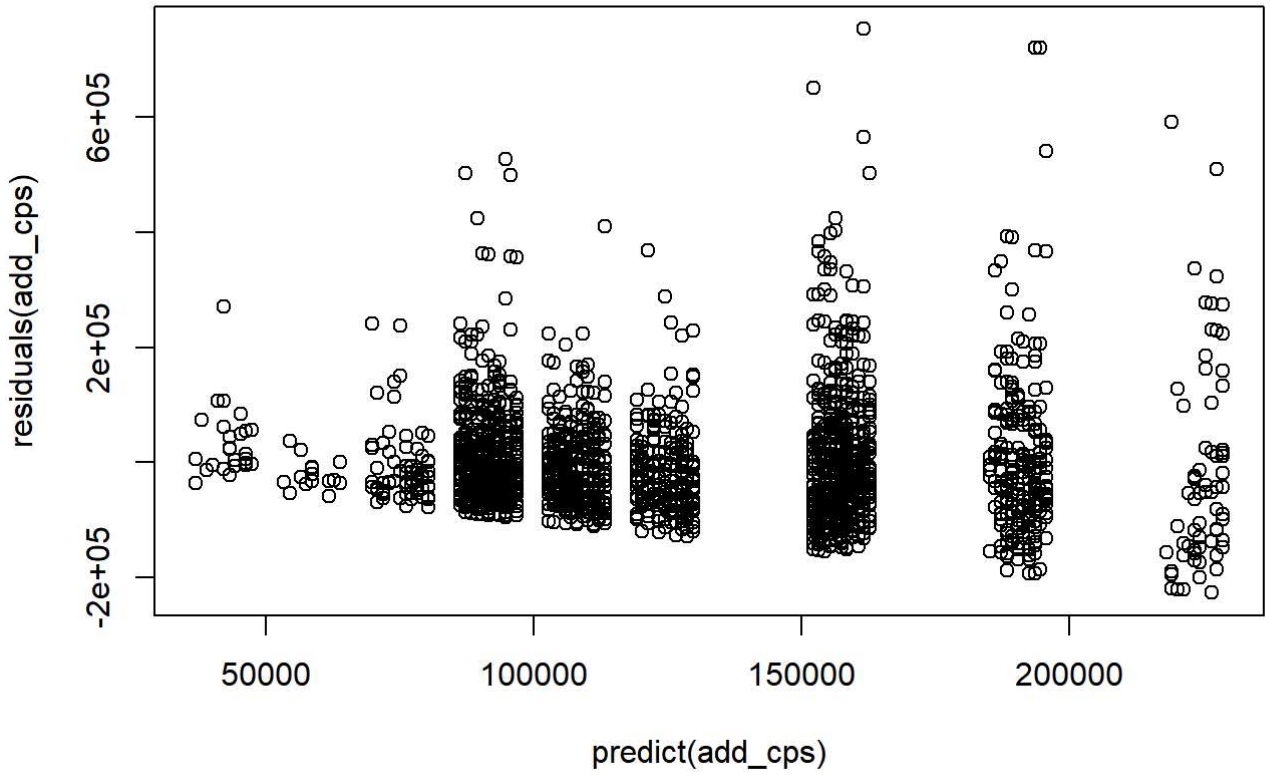
```
cps <- read.csv("California_CPS_degrees.csv")

add_cps <- lm(income ~ education + age, data =cps)

stargazer(add_cps,
          covariate.labels = c("Education", "Age"),
          title = "Income for People Ages 20-35", type = "text")
```

```
##
## Income for People Ages 20-35
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                 income
## ------------------------------------------------
## Education                    16,479.000***
##                               (889.794)
##
## Age                           1,051.582
##                               (694.354)
##
## Constant                    -137,806.300***
##                              (24,024.530)
##
## ------------------------------------------------
## Observations                    2,246
## R2                              0.134
## Adjusted R2                     0.134
## Residual Std. Error   101,756.700 (df = 2243)
## F Statistic          174.231*** (df = 2; 2243)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
#The standard deviation for these values assumes homoskedasticity, which may not be true.
df.mat <- data.frame(constant=rep(1, length (cps$income)), education = cps$education,
                     age=cps$age)
X<- data.matrix(df.mat)
Y <- cps$income

betas <- solve(t(X)%*%X)%*%t(X)%*%Y
print(betas)
```

```
##                  [,1]
## constant  -137806.340
## education   16478.997
## age          1051.582
```

```
#These Betas match the results from the additive linear regression. This represents an alternate way to find the relationships between the covariates and the treatment.

predicted<- predict(add_cps)
residual_cps <- residuals(add_cps)

plot(predict(add_cps), residuals(add_cps))
```



The data seems to be heteroskedastic because the larger income values seem to have a greater variance. Furthermore, there are a greater number of data points in the middle, which is lowering the variance. A more accuraete estimate of the standard deviation can be found in a few different ways.

```
#The standard deviation for each beta value using sandwhich
sand <- sandwich(add_cps)
robust_se <- sqrt(diag(sand))
print(robust_se)
```

```
## (Intercept)   education         age
##  25825.2054   1057.6145    719.3256
```

```
#Standard deviation for each beta using matrix multiplication.
solve(t(X)%*%X)%*%t(X)%*%diag(as.vector((as.vector(Y)-X%*%as.vector(betas)))*as.vector((as.vecto
r(Y)-X%*%as.vector(betas))))%*%X%*%solve(t(X)%*%X)
```

```
##            constant     education           age
## constant  666941234 -15029237.613 -15481152.468
## education -15029238   1118548.385      2254.142
## age       -15481152      2254.142    517429.329
```

```
matrix_sand<- solve(t(X)%*%X)%*%t(X)%*%diag(as.vector((as.vector(Y)-X%*%as.vector(betas)))*as.ve
ctor((as.vector(Y)-X%*%as.vector(betas))))%*%X%*%solve(t(X)%*%X)
print(sqrt(diag(matrix_sand)))
```

```
##   constant  education        age
## 25825.2054  1057.6145   719.3256
```

```
#Bootstrap for beta standard deviation
beta_bs = 0
beta_0 = 0
beta_1 = 0
beta_2 = 0

set.seed(6692)
for (i in 1:10000) {
  bs_ind <- sample(1:2246, replace=TRUE)
  df_bs <- cps[bs_ind, ]
  beta_bs <- lm(income ~ education + age, data=df_bs)
  beta_0[i] <- beta_bs$coefficient[1]
  beta_1[i] <- beta_bs$coefficient[2]
  beta_2[i] <- beta_bs$coefficient[3]
}

#New standard deviation calculation for divided by n, not n-1. Because the sample is so large, t
he new calculation is more standard.
standiv <- function(x) {
  mean <- mean(x)
  vector <- length(x)
  for(i in 1:length(x)){
    vector[i] <- (x[i]-mean)**2
  }
  division <- sum(vector)/length(x)
  answer <- division**.5
  print (answer)
}

sd(beta_0)
```

```
## [1] 25710.18
```

```
sd(beta_1)
```

```
## [1] 1066.961
```

```
sd(beta_2)
```

```
## [1] 715.9032
```

Neto and Cox

```
nc<- read.csv("netocox.csv")

#Balance table for ethnic groups
treat <- subset(nc, RUNOFF == 1)
control <- subset(nc, RUNOFF == 0)

mean_treat <- mean(treat$ENETH)
mean_control <- mean(control$ENETH)

balance <- data.frame(
  Group = c("No_Runoff", "Runoff"),
  Ethnic = c(mean_control, mean_treat))

stargazer(balance,
          type = "text",
          title = "Comparison for Pre-treatment Covariates",
          digits = 2,
          summary = FALSE)
```

```
##
## Comparison for Pre-treatment Covariates
## ==================
##     Group   Ethnic
## ------------------
## 1 No_Runoff  1.52
## 2  Runoff    1.64
## ------------------
```

```
#Additive linear regression
neto_add<- lm(ENPRES ~ RUNOFF + ENETH, data=nc)
summary(neto_add)
```

```
##
## Call:
## lm(formula = ENPRES ~ RUNOFF + ENETH, data = nc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3094 -1.0066 -0.2142  0.9920  1.9826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2629     0.8691   2.604   0.0219 *
## RUNOFF        0.6311     0.6087   1.037   0.3187
## ENETH         0.3661     0.4986   0.734   0.4759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.211 on 13 degrees of freedom
## Multiple R-squared:  0.1206, Adjusted R-squared:  -0.01464
## F-statistic: 0.8918 on 2 and 13 DF,  p-value: 0.4336
```

```
#Calculating APD
neto<- nc
neto$RUNOFF <- ifelse(neto$RUNOFF ==1, 0, 1)
neto$Additive_Predicted <- predict(neto_add, newdata = neto)

neto_inverse <- data.frame(
  "Country" = nc$COUNTRY,
  "Runoff" = nc$RUNOFF
)

neto_inverse$Y0 <- ifelse(neto$RUNOFF == 0, neto$Additive_Predicted, neto$ENPRES)
neto_inverse$Y1 <- ifelse(neto$RUNOFF ==0, neto$ENPRES, neto$Additive_Predicted)

neto_inverse$Treatment_Effect_Add <- neto_inverse$Y1 - neto_inverse$Y0

mean(neto_inverse$Treatment_Effect_Add)
```

```
## [1] 0.6310652
```

```
#APD = 0.631
print(neto_add)
```

```
##
## Call:
## lm(formula = ENPRES ~ RUNOFF + ENETH, data = nc)
##
## Coefficients:
## (Intercept)       RUNOFF        ENETH
##      2.2629       0.6311       0.3661
```

The calculated APD for the additive model is 0.631. Alternatively, the APD can be found from the partial derivative of the additive model (with respect to treatment). In this case, this is Beta 1 = 0.631.

```
#APD from interactive model
lm(ENPRES ~ RUNOFF + ENETH + RUNOFF:ENETH, data=nc)
```

```
##
## Call:
## lm(formula = ENPRES ~ RUNOFF + ENETH + RUNOFF:ENETH, data = nc)
##
## Coefficients:
##  (Intercept)        RUNOFF         ENETH  RUNOFF:ENETH
##       4.3034       -2.4911       -0.9792        2.0054
```

```
neto_int <- lm(ENPRES ~ RUNOFF + ENETH + RUNOFF:ENETH, data=nc)
summary(neto_int)
```

```
##
## Call:
## lm(formula = ENPRES ~ RUNOFF + ENETH + RUNOFF:ENETH, data = nc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8837 -0.6481  0.1422  0.4019  1.8485
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.3034     1.2285   3.503  0.00436 **
## RUNOFF        -2.4911     1.5605  -1.596  0.13640
## ENETH         -0.9792     0.7703  -1.271  0.22775
## RUNOFF:ENETH   2.0054     0.9405   2.132  0.05434 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 12 degrees of freedom
## Multiple R-squared:  0.3623, Adjusted R-squared:  0.2028
## F-statistic: 2.272 on 3 and 12 DF,  p-value: 0.1324
```

```
neto<- nc
neto$RUNOFF <- ifelse(neto$RUNOFF ==1, 0, 1)
neto$Interactive_Predicted <- predict(neto_int, newdata = neto)

neto_inverse <- data.frame(
  "Country" = nc$COUNTRY,
  "Runoff" = nc$RUNOFF
)

neto_inverse$Y0 <- ifelse(neto$RUNOFF == 0, neto$Interactive_Predicted, neto$ENPRES)
neto_inverse$Y1 <- ifelse(neto$RUNOFF ==0, neto$ENPRES, neto$Interactive_Predicted)

neto_inverse$Treatment_Effect <- neto_inverse$Y1 - neto_inverse$Y0
mean(neto_inverse$Treatment_Effect)
```

```
## [1] 0.6728034
```

```
#Alternative method:
#APD = B1 + B3(ENETH)
apd_int <- -2.4911 + 2.0054*nc$ENETH
mean(apd_int)
```

```
## [1] 0.6728369
```

```
#APD = 0.673

#Further treatment effect calculations:
ATE <- mean(neto_inverse$Treatment_Effect)
ATT <- mean(neto_inverse$Treatment_Effect[neto_inverse$Runoff ==1])
ATC <- mean(neto_inverse$Treatment_Effect[neto_inverse$Runoff==0])

print(ATE)
```

```
## [1] 0.6728034
```

```
print(ATT)
```

```
## [1] 0.7949634
```

```
print(ATC)
```

```
## [1] 0.5506433
```

```r
#APD for interactive polynomial model
nc$ENETH_SQ <- nc$ENETH**2
poly <- lm(ENPRES ~ RUNOFF + ENETH + RUNOFF * ENETH + ENETH_SQ + RUNOFF*ENETH_SQ, data = nc)

neto<- nc
neto$RUNOFF <- ifelse(neto$RUNOFF ==1, 0, 1)
neto$Poly_Predicted <- predict(poly, newdata = neto)

neto_poly <- data.frame(
  "Country" = nc$COUNTRY,
  "Runoff" = nc$RUNOFF)

neto_poly$Y0 <- ifelse(neto$RUNOFF == 0, neto$Poly_Predicted, neto$ENPRES)
neto_poly$Y1 <- ifelse(neto$RUNOFF == 0, neto$ENPRES, neto$Poly_Predicted)
neto_poly$Treatment_Effect <- neto_poly$Y1 - neto_poly$Y0

mean(neto_poly$Treatment_Effect)
```

```
## [1] 0.9913485
```

```r
#APD (with respect to RUNOFF) = B1 + B3(ENETH) + B5(ENETH^2)
print(poly)
```

```
##
## Call:
## lm(formula = ENPRES ~ RUNOFF + ENETH + RUNOFF * ENETH + ENETH_SQ +
##     RUNOFF * ENETH_SQ, data = nc)
##
## Coefficients:
##     (Intercept)           RUNOFF            ENETH         ENETH_SQ
##           7.510          -19.568           -5.103            1.198
##    RUNOFF:ENETH  RUNOFF:ENETH_SQ
##          23.950           -6.019
```

```r
apd_poly <- -19.568 + 23.950*nc$ENETH + -6.019*nc$ENETH_SQ
mean(apd_poly)
```

```
## [1] 0.9934198
```

```r
#APD = 0.991
```

Gerber Voter Experiment

```
gerber <- read.csv('GerberGreenLarimer_APSR_2008_social_pressure.csv')

gerber <- gerber %>% mutate(female = ifelse(sex == "male", 0, 1), age = 2006-yob, treated= ifels
e(gerber$treatment!=" Control", 1, 0), outcome = ifelse(gerber$voted=="Yes", 1, 0))

#Balance Table for Treatment versus Control
trt <- subset(gerber, treated == 1)
con <- subset(gerber, treated == 0)
mean_treated_num <- sapply(trt[, c("female", "age", "hh_size")], mean)
mean_control_num <- sapply(con[, c("female", "age", "hh_size")], mean)

mean_treated_bin <- sapply(trt[, c("g2000", "g2002", "g2004", "p2000", "p2002", "p2004")], funct
ion(x) mean(x %in% c("yes", "Yes")))
mean_control_bin <- sapply(con[, c("g2000", "g2002", "g2004", "p2000", "p2002", "p2004")], funct
ion(x) mean(x %in% c("yes", "Yes")))

comparison_df1 <- data.frame(
  Group = c("Treated", "Control"),
  Female = c(mean_treated_num['female'], mean_control_num['female']),
  Age = c(mean_treated_num['age'], mean_control_num['age']),
  Household_size = c(mean_treated_num['hh_size'], mean_control_num['hh_size']),
  General2000 = c(mean_treated_bin['g2000'], mean_control_bin['g2000']),
  General2002 = c(mean_treated_bin['g2002'], mean_control_bin['g2002']),
  Primary2000 = c(mean_treated_bin['p2000'], mean_control_bin['p2000']),
  Primary2002 = c(mean_treated_bin['p2002'], mean_control_bin['p2002']),
  Primary2004 = c(mean_treated_bin['p2004'], mean_control_bin['p2004']),
  N = c(nrow(trt),nrow(con))
)
stargazer(t(comparison_df1),
          type = "text",
          title = "Comparison for Pre-treatment Covariates",
          digits = 2,
          summary = FALSE)
```

```
##
## Comparison for Pre-treatment Covariates
## ================================
## Group           Treated   Control
## Female         0.4997088 0.4989411
## Age            49.75231  49.81355
## Household_size 2.184460  2.183667
## General2000    0.8420450 0.8433773
## General2002    0.8117194 0.8108950
## Primary2000    0.2515621 0.2518837
## Primary2002    0.3904188 0.3893737
## Primary2004    0.4029547 0.4003388
## N               152841    191243
## --------------------------------
```

```
#Balance Table for Treatment Type Assignment
civic <- subset(gerber, treatment == " Civic Duty")
haw <- subset(gerber, treatment == " Hawthorne")
self <- subset(gerber, treatment == " Self")
neib <- subset(gerber, treatment == " Neighbors")

mean_civic_num <- sapply(civic[, c("female", "age", "hh_size")], mean)
mean_haw_num <- sapply(haw[, c("female", "age", "hh_size")], mean)
mean_self_num <- sapply(self[, c("female", "age", "hh_size")], mean)
mean_neib_num <- sapply(neib[, c("female", "age", "hh_size")], mean)

mean_civic_bin <- sapply(civic[, c("g2000", "g2002", "g2004", "p2000", "p2002", "p2004")], funct
ion(x) mean(x %in% c("yes", "Yes")))
mean_haw_bin <- sapply(haw[, c("g2000", "g2002", "g2004", "p2000", "p2002", "p2004")], function
(x) mean(x %in% c("yes", "Yes")))
mean_self_bin <- sapply(self[, c("g2000", "g2002", "g2004", "p2000", "p2002", "p2004")], functio
n(x) mean(x %in% c("yes", "Yes")))
mean_neib_bin <- sapply(neib[, c("g2000", "g2002", "g2004", "p2000", "p2002", "p2004")], functio
n(x) mean(x %in% c("yes", "Yes")))

comparison_df2 <- data.frame(
  Group = c("Civic Duty", "Hawthorne", "Self", "Neighbors", "Control"),
  Female = c(mean_civic_num['female'], mean_haw_num['female'], mean_self_num['female'], mean_nei
b_num['female'],mean_control_num['female']),
  Age = c(mean_civic_num['age'], mean_haw_num['age'], mean_self_num['age'], mean_neib_num['ag
e'], mean_control_num['age']),
  Household_size = c(mean_civic_num['hh_size'], mean_haw_num['hh_size'], mean_self_num['hh_siz
e'], mean_neib_num['hh_size'], mean_control_num['hh_size']),
  General2000 = c(mean_civic_bin['g2000'], mean_haw_bin['g2000'], mean_self_bin['g2000'], mean_n
eib_bin['g2000'], mean_control_bin['g2000']),
  General2002 = c(mean_civic_bin['g2002'], mean_haw_bin['g2002'], mean_self_bin['g2002'], mean_n
eib_bin['g2002'], mean_control_bin['g2002']),
  Primary2000 = c(mean_civic_bin['p2000'], mean_haw_bin['p2000'], mean_self_bin['p2000'], mean_n
eib_bin['p2000'], mean_control_bin['p2000']),
  Primary2002 = c(mean_civic_bin['p2002'], mean_haw_bin['p2002'], mean_self_bin['p2002'], mean_n
eib_bin['p2002'], mean_control_bin['p2002']),
  Primary2004 = c(mean_civic_bin['p2004'], mean_haw_bin['p2004'], mean_self_bin['p2004'], mean_n
eib_bin['p2004'], mean_control_bin['p2004']),
  N = c(nrow(civic), nrow(haw), nrow(self), nrow(neib), nrow(con))
)
stargazer(t(comparison_df2),
          type = "text",
          title = "Comparison for Pre-treatment Covariates",
          digits = 2,
          summary = FALSE)
```

```
##
## Comparison for Pre-treatment Covariates
## ================================================================
## Group          Civic Duty Hawthorne   Self    Neighbors  Control
## Female         0.5001832  0.4990053 0.4995813 0.5000654 0.4989411
## Age             49.65904  49.70480  49.79251  49.85294  49.81355
## Household_size  2.189126  2.180138  2.180805  2.187770  2.183667
## General2000    0.8417238  0.8444142 0.8403893 0.8416534 0.8433773
## General2002    0.8111099  0.8129515 0.8114763 0.8113400 0.8108950
## Primary2000    0.2535716  0.2503665 0.2511120 0.2511976 0.2518837
## Primary2002    0.3888482  0.3943304 0.3919096 0.3865867 0.3893737
## Primary2004    0.3994453  0.4032300 0.4024805 0.4066647 0.4003388
## N                38218      38204     38218     38201    191243
## ----------------------------------------------------------------
```

```
#Voter Turnout for each treatment
civic <- civic %>%
  mutate(osix = ifelse(voted=="Yes", 1, 0))

haw <- haw %>%
  mutate(osix = ifelse(voted== "Yes", 1, 0))

self <- self %>%
  mutate(osix = ifelse(voted== "Yes", 1, 0))

neib <- neib %>%
  mutate(osix = ifelse(voted== "Yes", 1, 0))

con <- con %>%
  mutate(osix = ifelse(voted== "Yes", 1, 0))

means_df2 <- data.frame(
  Group = c("Civic Duty", "Hawthorne", "Self", "Neighbors", "Control"),
  Voter_Turnout = c(mean(civic$osix), mean(haw$osix), mean(self$osix), mean(neib$osix), mean(con
$osix))
)
stargazer(t(means_df2),
          type = "text",
          title = "2006 Primary Turnout for Covariates",
          digits = 2,
          summary = FALSE)
```

```
##
## 2006 Primary Turnout for Covariates
## ================================================================
## Group         Civic Duty Hawthorne   Self    Neighbors  Control
## Voter_Turnout 0.3145377  0.3223746 0.3451515 0.3779482 0.2966383
## ----------------------------------------------------------------
```

```
#Treatment effect for each treatment
gerber <- gerber %>% mutate(civic = ifelse(treatment == " Civic Duty", 1, 0), haw = ifelse(treat
ment == " Hawthorne", 1, 0), self = ifelse(treatment == " Self", 1, 0), neib = ifelse(treatment
== " Neighbors", 1, 0))

mu_hat_control <- mean(gerber$outcome[gerber$treated==0])
mu_hat_civic <- mean(gerber$outcome[gerber$civic==1])
mu_hat_haw <- mean(gerber$outcome[gerber$haw==1])
mu_hat_self <- mean(gerber$outcome[gerber$self==1])
mu_hat_neib <- mean(gerber$outcome[gerber$neib==1])

mu_hat_civic - mu_hat_control
```

```
## [1] 0.01789934
```

```
mu_hat_haw - mu_hat_civic
```

```
## [1] 0.007836968
```

```
mu_hat_self - mu_hat_haw
```

```
## [1] 0.02277688
```

```
mu_hat_neib - mu_hat_self
```

```
## [1] 0.03279672
```

The civic treatment effect is a 1.79% increase from no treatment in voter turnout for the 2006 election. The hawthorne treatment effect is a .78% increase from civic treatment in voter turnout for the 2006 election. The self treatment effect is a 2.28% increase from the hawthorne treatment in voter turnout for the 2006 election. The neighbor treatment effect is a 3.28% increase from the self treatment in voter turnout for the 2006 election.

```
#Bootstrapped treatment effect
set.seed(9090)
n_ite <- 1000
tau_neib <- rep(NA, n_ite)
tau_self <- rep(NA, n_ite)
tau_haw <- rep(NA, n_ite)
tau_civic <- rep(NA, n_ite)
n_sample <- nrow(gerber)

for(i in 1:n_ite){
  index_bs <- sample(1:n_sample, replace = T)
  df_bs <- gerber[index_bs, ]
  tau_neib[i] <- mean(df_bs$outcome[df_bs$neib==1])-mean(df_bs$outcome[df_bs$self==1])
  tau_self[i] <- mean(df_bs$outcome[df_bs$self==1])-mean(df_bs$outcome[df_bs$haw==1])
  tau_haw[i] <- mean(df_bs$outcome[df_bs$haw==1])-mean(df_bs$outcome[df_bs$civic==1])
  tau_civic[i] <- mean(df_bs$outcome[df_bs$civic==1])-mean(df_bs$outcome[df_bs$treated==0])
}
tau_quantiles <- t(sapply(list(tau_neib, tau_self, tau_haw, tau_civic), function(x) quantile(x,
c(0.025, 0.975))))
bootstrap_df <- data.frame(tau_quantiles[,1], tau_quantiles[,2],sapply(list(tau_neib, tau_self,
tau_haw, tau_civic),mean))
colnames(bootstrap_df) <- c("Lower Bound", "Upper Bound", "Mean")
rownames(bootstrap_df) <- c("Neighbors", "Self", "Hawthorne", "Civic")
bootstrap_df
```

```
##           Lower Bound Upper Bound        Mean
## Neighbors 0.025824664  0.03932459 0.032853901
## Self      0.016004182  0.02958043 0.022757738
## Hawthorne 0.001210968  0.01404907 0.007757274
## Civic     0.012799730  0.02289875 0.017965335
```

```
#Additive linear regression
gerber <- gerber%>%mutate(civic_num = ifelse(gerber$treatment == " Civic Duty", 1,0),
                          haw_num = ifelse(gerber$treatment == " Hawthorne", 1,0),
                          self_num = ifelse(gerber$treatment == " Self", 1,0),
                          neib_num = ifelse(gerber$treatment == " Neighbors", 1,0))

gerber <- gerber %>%
  mutate(osix = ifelse(voted== "Yes", 1, 0))

lm1 <- lm(osix ~ civic_num + haw_num + self_num + neib_num, data = gerber)

stargazer(lm1,
          covariate.labels = c("Civic", "Hawthorne", "Self", "Neighbor"),
          title = "2006 Voting", type = "text")
```

```
##
## 2006 Voting
## ================================================
##                           Dependent variable:
##                       ---------------------------
##                                   osix
## ------------------------------------------------
## Civic                           0.018***
##                                  (0.003)
##
## Hawthorne                       0.026***
##                                  (0.003)
##
## Self                            0.049***
##                                  (0.003)
##
## Neighbor                        0.081***
##                                  (0.003)
##
## Constant                        0.297***
##                                  (0.001)
##
## ------------------------------------------------
## Observations                     344,084
## R2                                0.003
## Adjusted R2                       0.003
## Residual Std. Error      0.464 (df = 344079)
## F Statistic          292.976*** (df = 4; 344079)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The typical American not receiving a treatment will vote in the 2006 election 29.7% of the time (or more accurately, 29.7% of Americans will vote without a treatment). This percent increases by 1.8 with the Civic treatment, by 2.6 with the Hawthorne treatment, by 4.9 with the self treatment, or 8.1 with the neighbor treatment (from the control).

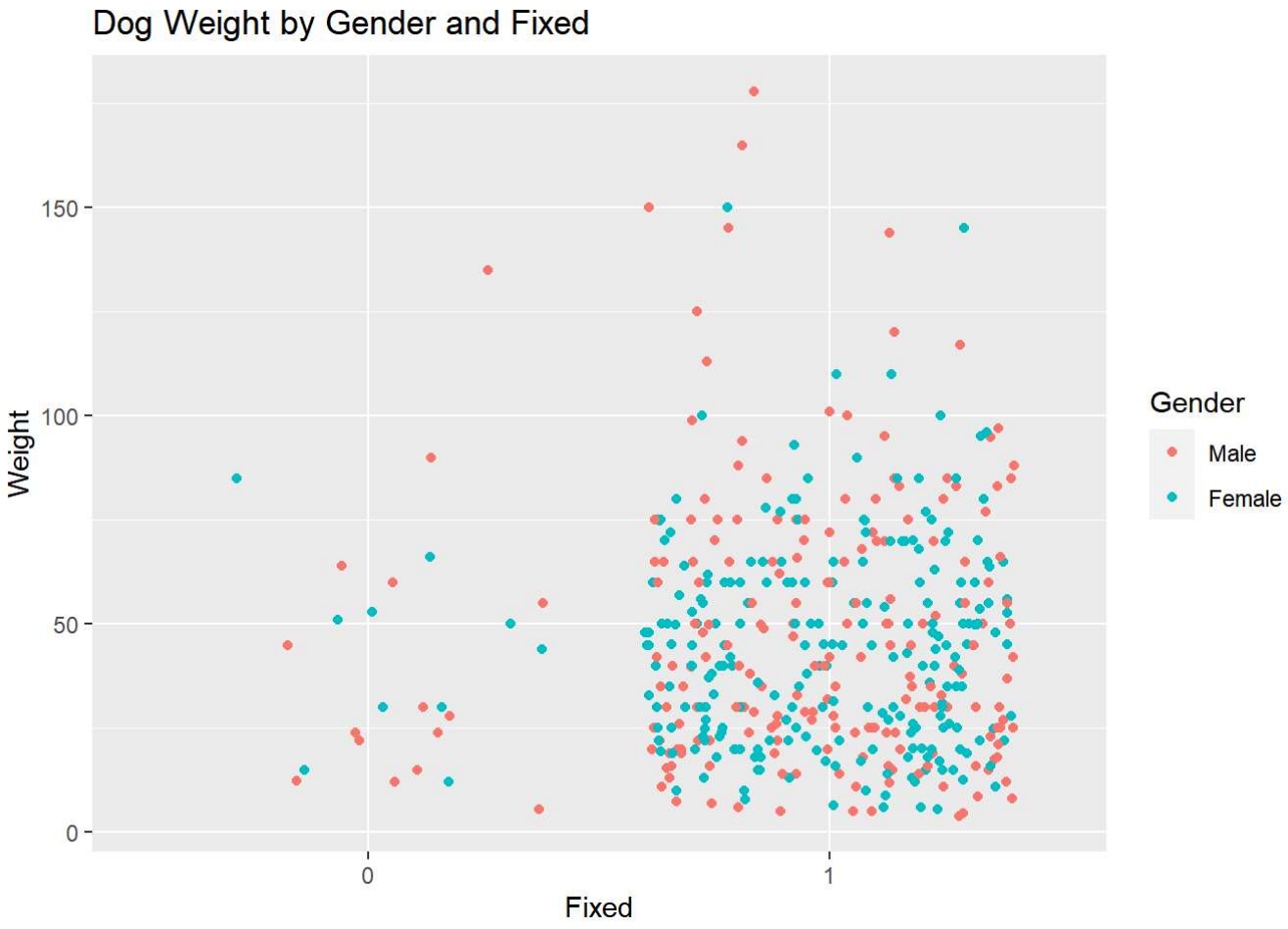
```
#Expanded additive linear regression
gerber <- gerber%>%mutate(p2000_num = ifelse(gerber$p2000 == "yes", 1,0),
                          p2002_num = ifelse(gerber$p2002 == "yes", 1,0),
                          p2004_num = ifelse(gerber$p2004 == "Yes", 1,0),
                          g2000_num = ifelse(gerber$g2000 == "yes", 1,0),
                          g2002_num = ifelse(gerber$g2002 == "yes", 1,0))

lm2 <- lm(osix ~ civic_num + haw_num + self_num + neib_num + p2000_num + p2002_num + p2004_num +
g2000_num + g2002_num, data = gerber)

stargazer(lm1, lm2,
          covariate.labels = c("Civic", "Hawthorne", "Self", "Neighbor", "2000 Primary", "2004 P
rimary", "2002 Primary", "General 2000", "General 2002"),
          title = "2006 Voting", type = "text")
```

```
##
## 2006 Voting
## ================================================================================
##                                        Dependent variable:
##                    ------------------------------------------------------------
##                                              osix
##                                (1)                           (2)
## --------------------------------------------------------------------------------
## Civic                        0.018***                      0.018***
##                               (0.003)                       (0.003)
##
## Hawthorne                    0.026***                      0.025***
##                               (0.003)                       (0.003)
##
## Self                         0.049***                      0.048***
##                               (0.003)                       (0.003)
##
## Neighbor                     0.081***                      0.081***
##                               (0.003)                       (0.003)
##
## 2000 Primary                                               0.100***
##                                                             (0.002)
##
## 2004 Primary                                               0.133***
##                                                             (0.002)
##
## 2002 Primary                                               0.156***
##                                                             (0.002)
##
## General 2000                                                -0.003
##                                                             (0.002)
##
## General 2002                                               0.101***
##                                                             (0.002)
##
## Constant                     0.297***                      0.077***
##                               (0.001)                       (0.002)
##
## --------------------------------------------------------------------------------
## Observations                  344,084                       344,084
## R2                             0.003                         0.075
## Adjusted R2                    0.003                         0.075
## Residual Std. Error     0.464 (df = 344079)          0.447 (df = 344074)
## F Statistic         292.976*** (df = 4; 344079) 3,101.000*** (df = 9; 344074)
## ================================================================================
## Note:                                            *p<0.1; **p<0.05; ***p<0.01
```

The values are similar between the two regressions. There is only a minute difference between the two regression for the Hawthorne and Self Treatments.

Dogs

```
dogs <- read.csv('dogs.csv')

#Graph of weight based on gender and fixed status
dogs$Gender[dogs$Gender == "Male"] <- 0
dogs$Gender[dogs$Gender == "Female"] <- 1

dogs$Fixed[dogs$Fixed=='No'] <- 0
dogs$Fixed[dogs$Fixed=='Yes'] <- 1
dogs$Fixed[dogs$Fixed=='']<- NA

dogs %>%
  drop_na() %>%
  ggplot(aes(x=Fixed, y= Weight))+
  geom_point(aes(color=Gender), position='jitter')+
  labs(title= 'Dog Weight by Gender and Fixed') +
  scale_color_discrete(labels=c('Male', 'Female'))
```



```
#Linear Regressions
lm_add <- lm(Weight~Gender + Fixed, data=dogs)

lm_gender <- lm(Weight ~ Gender, data=dogs)
lm_fixed <- lm(Weight~ Fixed, data=dogs)
lm_interactive <- lm(Weight~Gender + Fixed+ Gender*Fixed, data=dogs)

stargazer(lm_add, lm_interactive,
          covariate.labels = c("Gender", "Fixed", "Gender:Fixed"),
          title = "Dogs Weight", type = "text")
```

```
##
## Dogs Weight
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                       Weight
##                               (1)                 (2)
## ---------------------------------------------------------------
## Gender                      -2.659               2.152
##                             (2.731)            (11.733)
##
## Fixed                        3.786               5.863
##                             (5.925)             (7.710)
##
## Gender:Fixed                                    -5.086
##                                                (12.065)
##
## Constant                   43.397***           41.473***
##                             (5.846)             (7.421)
##
## ---------------------------------------------------------------
## Observations                  448                 448
## R2                           0.003               0.003
## Adjusted R2                 -0.002              -0.003
## Residual Std. Error  28.714 (df = 445)   28.741 (df = 444)
## F Statistic         0.637 (df = 2; 445) 0.483 (df = 3; 444)
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

```
#Comparison with sample data means
mean(dogs$Weight[dogs$Gender==1 & dogs$Fixed==1], na.rm=TRUE)
```

```
## [1] 44.40197
```

```
mean(dogs$Weight[dogs$Gender==1 & dogs$Fixed==0], na.rm=TRUE)
```

```
## [1] 43.625
```

```
mean(dogs$Weight[dogs$Gender==0 & dogs$Fixed==1], na.rm=TRUE)
```

```
## [1] 47.33651
```

```
mean(dogs$Weight[dogs$Gender==0 & dogs$Fixed==0], na.rm=TRUE)
```

```
## [1] 41.47333
```